#2
2-13-01

501.38779X00

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant(s): Toru HISAMITSU, et al

Serial No.: 09/642,771

Filed: August 22, 2000

Title: WORD IMPORTANCE CALCULATION METHOD, DOCUMENT RETRIEVING INTERFACE, WORD DICTIONARY MAKING METHOD

Group: 2175

LETTER CLAIMING RIGHT OF PRIORITY

Honorable Commissioner of
Patents and Trademarks
Washington, D.C. 20231

August 22, 2000

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, the applicant(s) hereby claim(s) the right of priority based on Japanese Patent Application No.(s) 11-237845 filed August 25, 1999.

A certified copy of said Japanese Application is attached.

Respectfully submitted,

ANTONELLI, TERRY, STOUT & KRAUS, LLP

Carl I. Brundidge
Registration No. 29,621

CIB/nac
Attachment
(703) 312-6600

3c990c743

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年　８月２５日

出 願 番 号
Application Number: 平成１１年特許願第２３７８４５号

出 願 人
Applicant (s): 株式会社日立製作所

２０００年　４月２１日

特 許 庁 長 官
Commissioner,
Patent Office 近 藤 隆 彦

出証番号　出証特２０００－３０２８６０８

【書類名】　特許願

【整理番号】　H99017431A

【提出日】　平成11年 8月25日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 17/30

【発明者】

【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内

【氏名】　久光　徹

【発明者】

【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地
株式会社日立製作所中央研究所内

【氏名】　丹羽　芳樹

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社日立製作所

【代理人】

【識別番号】　100075096

【弁理士】

【氏名又は名称】　作田　康夫

【電話番号】　03-3212-1111

【手数料の表示】

【予納台帳番号】　013088

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　単語の重要度計算方法、文献検索インタフェイス、単語辞書作成方法

【特許請求の範囲】

【請求項１】

文書集合に含まれる単語の重要度を計算する方法において、抽出すべき単語を含む部分文書集合と、全文書集合との乖離度を用いて該抽出すべき単語の重要度を計算することを特徴とする単語の重要度計算方法。

【請求項２】

請求項１記載の単語の重要度計算方法において、

上記乖離度は、上記部分集合と、上記全文書集合との距離ｄと、上記部分集合と同程度の単語数を含み、かつ、上記全文書集合からランダム選出された部分文書集合と、上記全文書集合との距離ｄ’あるいはは該ｄ’の推定値とを比較して求めることを特徴とする単語の重要度計算方法。

【請求項３】

請求項２記載の単語の重要度計算方法において、

二つの文書集合間の距離ｄは、それぞれの文書集合における単語分布、すなわち全文書集合に含まれる各単語の出現確率を用いて計算することを特徴とする単語の重要度計算方法。

【請求項４】

請求項２または請求項３のいずれかに記載の単語の重要度計算方法において、

上記抽出すべき単語を含む部分文書集合に含まれる文書数が所定数より大きい場合、該部分文書集合からランダムサンプリングによりあらかじめ定めた数の文書を抽出し、該抽出された文書の集合と上記全文書集合との乖離度を用いて、該部分文書集合と該全文書集合の乖離度を推定することを特徴とする単語の重要度計算方法。

【請求項５】

文書集合を特徴付ける単語を画面に提示する機能を持つ文献検索インタフェイスにおいて、全文書集合中に現れる各単語について、該単語を含む部分文書集合

中の単語分布と、該全文書集合中の単語分布との乖離度を用いて各々の単語の重要度を計算し、該重要度を画面に提示する単語の選択、配置、または、配色に反映させることを特徴とする文献検索インタフェイス。

【請求項６】

文書集合を特徴付ける単語を画面に提示する機能を持つ文献検索インタフェイスにおいて、検索の結果得られる文書集合中に現れる各単語について、該単語を含む該検索の結果得られる文書集合の部分文書集合中の単語分布と、該検索の結果得られる文書集合中の単語分布との乖離度を用いて各々の単語の重要度を計算し、該重要度を画面に提示する単語の選択、配置、または、配色に反映させてることを特徴とする文献検索インタフェイス。

【請求項７】

文書集合からあらかじめ与えた規則に従い重要単語を抽出する単語辞書作成方法において、全文書集合中に現れる各単語について、該単語を含む部分文書集合と、該全文書集合との乖離度を用いて各々の単語の重要度を計算し、該重要度に基づいて抽出すべき単語を選択することを特徴とする単語辞書作成方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、文書群中の単語または単語列の重要性を測る技術に係り、文献検索の支援、単語辞書の自動作成等に利用される。

【０００２】

【従来の技術】

図１は、検索された文書内の「特徴単語」を提示するウィンドウを持つ文書検索システムの例であるが、右側のウィンドウには左側に示されている文書中の単語が選択されて表示されている。このような検索システムの例として、例えば、特開平10-74210「文献検索支援方法及び装置およびこれを用いた文献検索サービス」(文献１)があげられる。

【０００３】

また、影浦峡（他）、"Methodsof automatic term recognition: A review"、

Terminology、 1998)(文献２)には、単語の重要度を計算する方法が記載されている。単語の重要度を計算する方法は、専門用語の自動抽出や、文献検索の際に文書を特徴付ける単語に重みをつけることを目的として、長い間研究されてきた。

【０００４】

単語の重み付けに関しては、特定の文書内から重要語を抽出することを目的とするもの、全文書から重要語を抽出するのを目的とするものがある。前者についてもっとも有名なものは、tf-idfである。idfは、全文書数Nをある単語wが現れる文書数N(w)で割ったものの対数、tfは単語の文書d内での出現頻度f(w、D)であり、tf-idfは、これらの積として

f(w、d) × log2(N/N(w)) で表される。次のような、f(w、d)の平方根をとる等の変形がある

f(w、d)**0。５ × log2(N/N(w))他にもさまざまな変形があるが、tf-idfの基本的な性質として、「単語がより多く、より少ない文書に偏って出現するほど大きくなる」ように設定される。

【０００５】

文献２には記述されていないが、この指標を特定の文書中での単語の重要度でなく、文書集合全体での単語の重要度を測る指標に拡張する自然な方法は、f(w、d)を、wの全文書中での頻度f(w)に置き換えることである。

【０００６】

全文書中から重要語を抽出するための方法の一つとして、単語の出現の偏りをより精密に捕えるために、注目する単語の、与えられた文書カテゴリごとの出現頻度の差異の偶然性を測り、偶然でない度合いが高いものを重要語としようという方法があり、尺度としてχ2検定等が利用されているが、この場合、文書集合はあらかじめカテゴリに分類されている必要がある。

【０００７】

これらとは別系統の研究として、自然言語処理の立場から重要語としてふさわしい語のまとまりを捕えようとする一連の研究がある。これらの研究においては、文法知識を用いて語の並びに制約を加えるとともに、隣り合う単語の共起の強

さをさまざまな尺度で測る方法が提案されている。尺度としては、(各点)相互情報量、対数尤度比等が利用されている。

【０００８】

【発明が解決しようとする課題】

従来用いられてきた手法には、以下の問題があった：(1) tf-idf(もしくはその類似手法)の精度は不充分である。経験上語の頻度の寄与が大きすぎる傾向があり、例えば「する」のような一般的すぎる不用語の排除ができにくい。(2) 特定の語のカテゴリ間での分布の違いを比較する方法では、あらかじめ文書が分類されている必要があるが、この条件は一般に満たされない。(3) 隣り合う単語の共起の強さを利用する手法では、１単語のみの場合重要度が評価できない。n>2について、拡張が自明ではない。(4) 従来は重要／非重要を分ける閾値の設定が困難かつ恣意的になりがちであった。本発明の目的は、このような問題の無い方法を提案することである。

【０００９】

【課題を解決するための手段】

以下、タームとは、単語または単語列のことを示す。「タームの重要さ」を、専門用語抽出や、情報検索の観点から言い替えると、あるタームが重要であるとは、そのタームがある程度のまとまった話題を想起させる、すなわち informative または domain-specific であることと言える。これは情報検索の領域ではrepresentative と呼ばれ、この意味でのタームの「重要性」は、representativeness とも呼ばれる。このようなタームは、文書集合の内容を俯瞰するときに役立つと考えられるため、情報検索やその支援システムにおいては重要である。

【００１０】

representativenessを測る際、従来の方法は注目するターム自体の分布に着目していた。しかし、tf-idfのような手法は簡易ではあるが精度が不足し、χ2等の統計量を用いる手法では、一つのタームは例外を除きせいぜい数十回しか現れないため、多くのタームについて統計的に意味のある値を得るのは困難であり、それが精度の低下につながっていた。

【００１１】

本発明は、特定のタームの分布でなく、注目するタームと一緒にあらわれる単語の分布に注目する。これは、「タームの重要度は、そのタームといっしょに現れる単語の分布の偏りと関係がある」という作業仮説に立つものであり、あるタームが「重要である」ということを、「そのタームと共に現われる単語の単語分布が特徴的である」と解釈する。

【００１２】

そこで、本発明では、上記課題を解決するため、文書集合に含まれる単語の重要度を計算する際に、抽出すべき単語を含む部分文書集合中の単語分布と、もととなる文書集合中の単語分布との乖離度を用いるこをを特徴とする。特に、上記乖離度は、上記部分集合と上記文書集合との距離ｄと、上記部分集合と同程度の単語数を含み、かつ、上記文章集合からランダム選出された部分文章集合と上記文章集合との距離ｄ’とを比較して求めることを特徴とする。

【００１３】

【発明の実施の形態】

以下では、任意のタームのrepresentativenessを求める手法と、その情報検索システムへの応用を示す。まず、上記「課題を解決するための手段」の欄で述べた考え方を数学的に言い替えることにより、タームのrepresentativenessを測る指標を導入する。すなわち、任意のタームW(単語または単語列)について、Wを含む文書すべての集合における単語分布と、全文書の単語分布の距離に着目する。具体的には、W: ターム(任意の個数の単語からなる)、D(W): Wを含む文書すべての集合、D0 :全文書の集合、PD(W): D(W)における単語分布、P0 :D0 における単語分布、とするとき、Wのrepresentativeness Rep(W)を、２つの分布{PD(W)、 P0 }の距離 Dist{PD(W)、 P0 }に基づいて定義する。

【００１４】

単語分布間の距離の計測の方法としては、主要なものだけでも、(1) 対数尤度比(log-likelihood ratio)、(2) Kullback-Leibler divergence、(3) transition probability、(4) vector-space model (cosign 法)等が考えられるが、例えば対数尤度比を用いて安定した結果が得られることを確認している。全単語を{w

1、。。。。、wn}、ki とKi を、単語wi がD(W)、D0 に出現する頻度として、対数尤度比を用いた場合のPD(W) と P0 の距離を以下で定義する。

【００１５】

【数１】

数１

$$\sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} - \sum_{k=1}^{n} k_i \log \frac{K_i}{\#D_0}$$

【００１６】

図2は、日経新聞1996年版の記事を用い、そこにあらわれるいくつかの語Wに対し、各語Wについて、D(W)の含む単語数#D(W)を横軸に、Dist{PD(W)、 P0 }を縦軸にプロットしたものである。ここでは、距離として対数尤度比を用いている。図２から見られるとおり、#D(W)が近いターム同士で比較すれば、たとえば「米国」は「する」、「オウム」は「結び付ける」よりDist{PD(W)、 P0}の値が高く直感と合致する。しかし、このままでは#D(W)が離れたターム(これは概ね、二つのタームの頻度が大きく異なることとと等価である)同士のrepresentativenessを適切に比較することができない。なぜならば、一般にDist{PD(W)、 P0 }は、#D(W)が大きくなるにつれて増加するからである。実際、「オウム」は「する」とDist{PD(W)、 P0 }の値が同程度となる。 そこで特定のタームから離れてDist{・、P0}の振る舞いを調べるため、さまざまな数の文書をランダムサンプリングし、その結果得られたさまざまな大きさの文書集合Ｄに対して計算した(#D、 Dist{PD、 P0 })を、図２に「×」を用いてプロットした。これらの点は、(0、 0)に始まり(#D0、 0)に終わる一つのなめらかな曲線により良く近似できると思われる。以下、この曲線をベースライン曲線と呼ぶことにする。

【００１７】

D ＝φのときと、D ＝ D0のときにDist{PD、 P0 }が0となるのは定義から明らかであるが、 #D ＝ 0 付近の挙動は、比較的全文書数が少ないとき(2、000文書程度)から、新聞1年分(3000、000文書程度)まで、全文書集合が様々な大きさの

場合にかなり安定して近似できることが確認できた。

【００１８】

そこで、上記のさまざまな大きさの全文書集合において、ベースライン曲線が指数関数を用いた近似関数を用いて安定して精度良く求められる区間(1000 ≦ #D < 20000)上で近似関数B(･)を求め、1000 ≦ #D(W) < 20000を満たすWのrepresenentativeness を、Dist{PD(W)、 P0 }に、B(･)による正規化を施した値： Rep(W) = Dist{PD(W)、 P0 }/B(#D(W))により定義する（ただし、ここでいう単語は、記号や助詞、格助詞などの情報検索の検索語として確実に不要とみなされたものはすでに除いたものを指す。これらを含めた場合でも同様の手法が実現できるが、その場合は上記の数字は若干異なってくる）。

【００１９】

ここで、「する」のように著しく#D(W)が大きい場合でも、上記のベースライン関数の有効域を用いることを可能にすることと、計算量を低減することを意図して、20、000<#D(W)となるようなWに対しては、D(W)として150文書程度をランダム抽出し、1000 ≦ #D(W) < 20000を満たすようにしてからRep(W)を計算する。

【００２０】

一方、上記の区間で求めたベースライン曲線の近似関数は、{x | 0≦x<1000}で、値を大きめに見積もる傾向があるため、#D(W)≦1000となるWについては、正規化の結果Rep(W)は低めに出る。しかし、1000単語はほぼ新聞の２、３記事に相当するが、出現文書数がその程度のタームは我々の目的からの重要度は低いため、そのまま適用した。もちろん、別のベースラインを計算しておいてもよい。

ランダムサンプリングした文書集合DにおけるDist{PD、 P0 }/B(#D)は、さまざまなコーパスにおいて、安定して平均Avrがほぼ1(±0。01)、標準偏差σが0。05程度であった。また、最大値が Avr + 4σ を越えることはなかったので、あるタームWのRep(W)の値が、「意味のある値である」と判断するための閾値として、Avr + 4σ=1。20を設ける。

【００２１】

上記指標Rep(･)は、(1) 数学的な意味付けが明瞭であり、(2) 高頻度タームと

低頻度タームの比較が自然にできる。(3) 閾値の設定が自然にできる。(4) 任意の長さのタームに対して適用できる。等の好ましい性質を持つ。

【００２２】

本発明で提案する指標Rep(・)の有効性は、実験によっても確認されている。日経新聞1996年版の記事中、総頻度が3以上の単語から20、000語を無作為抽出し、そのうちの2、000個を、検索内容の概観に現われることが「好ましい a」「どちらでもよい」「好ましくない d」の3種類に人手で分類した。該20、000語を何らかの方法でソートしたときに、各クラスに分類された語の、先頭からN位までの累積出現頻度グラフを、ランダムソート、頻度、「従来の技術」において述べた、全文書を対象としたtf-idfの変形版、すなわち、 Nを全文書数、 N(w)をwが現れる文書数、f(w)をwの全文書中での頻度として、 f(w)**0。5 × log2(N/N(w))を用いた。

【００２３】

図７は、分類が"a"となったものの累積頻度を、ランダム、頻度、tf-idf、新指標のそれぞれを用いた場合で比較したものである。グラフから明らかに、ランダム＜頻度＜tf-idf＜新指標の順で「好ましい」と分類される語の優先順位を上げる力が強い。改善はあきらかに有為である。図８は、分類が"d"となったものの累積頻度の比較であり、新指標の選別能力の優位性がより際立っている。頻度とtf-idfはランダムな場合と変わらず、「不要語」特定能力の低さを現している。このため、本発明で提案する指標は、不要語の同定にとりわけ有効であり、「高頻度かつrepresentative-nessの低い語」を選ぶことによる、stop-word listの自動作成や、文献類似度計算における語の重み付けの精度改善等への応用が期待される。

【００２４】

これまでに述べたrepresentativeness を計算するためのシステム構成例を図３に示す。以下図３、４を用いてrepresentativeness の計算について述べる。３０１は記憶装置であり、ハードディスク等を用いて文書データ、各種のプログラム等を格納する。また、プログラムの作業用領域としても利用される。以下、３０１１は、文書データ。以下の例では日本語を用いるが、言語にはよらない。

３０１２は、形態素解析プログラム、文書を構成する単語を同定する。日本語の場合は分かち書き＋品詞付け、英語の場合は原型還元等の処理を行う。この手法については特定しない。両言語とも、商用・研究用をとわずさまざまなシステムが公開されている。３０１３は、単語・文書対応付けプログラム。形態素解析の結果から、どの単語がどの文書に何回あらわれているか、逆にどの文書にどのような単語が何回あらわれているかを調べる。基本的には単語と文書をそれぞれ行・列とする行列の要素を計数により埋める作業であり、この手法については特定しない。３０１４は、単語・文書対応データベース（ＤＢ）。上記で計算された単語・文書対応データを記録するＤＢ。３０１５は、representativeness計算プログラム。図４にその詳細を示す、タームのrepresentativenessを計算するプログラム。３０１６は、計算されたタームのrepresentativenessを記録するＤＢ。３０１７は、複数のプログラム間で共通に参照するデータを記録する領域である。３０１８は、作業用の領域である。３０２は、入力装置、３０３は、通信装置、３０４は、メインメモリ、３０５は、CPU、３０６は、ディスプレイ、キーボード等より構成される端末装置、である。

【００２５】

図４は、３０１５のrepresentativeness計算プログラムの詳細である。以下、これを用いて、特定のタームのrepresentativenessを求める手法を説明する。４０１１は、背景単語分布計算モジュールである。このモジュールは、一度だけ用いられ、各単語の全文中での頻度を記録する。すなわち、(数１)と同じく、全単語を{w1、。。。。、wn}とし、Ki を単語wi が全文書D0 中に出現する頻度として、(K1、 。。。。、 Kn)を記録する。４０１２は、与えられた文書データに対してベースライン関数を推定するモジュールである。このモジュールも、はじめに一度だけ用いられる。次の基本的な要素の組み合わせで実現できる。すなわち、

（１）文書集合が与えられた時、含まれる単語数が1000語前後から20、000語前後になるような文書集合をランダムに、しかも含まれる単語数が1000から20、000語の間でできるだけさまざまな値をとるように選び、それぞれの場合に(数１)を用いて４０１１で求めた背景単語分布との距離を計算する。

【００２６】

（２）（１）で得られた点群と最小２乗法等を用いてベースライン関数B(・)を算出する。B(・)は、単語数から正の実数への関数である。これらに関する方法は特定しない。標準的な手法が存在する。

【００２７】

４０１３は、部分文書集合抽出モジュールである。（「タームW＝wn1…wnkがあたえらたとき、単語・文書対応ＤＢ３０１４により、各単語wni(1≦i≦k)を含む文書集合D(wni)(1≦i≦k)を求め、すべてのD(wni)(1≦i≦k)の共通集合を取ってD(W)をもとめる。」→）タームW＝wn1…wnkがあたえらたとき、単語・文書対応ＤＢ３０１４により、各単語wni(1≦i≦k)を含む文書集合D(wni)(1≦i≦k)を求める。単語・文書対応ＤＢ３０１４に単語の文書内の位置情報まで記録されていると仮定すれば、タームW＝wn1…wnkを含む文書集合は、すべてのD(wni)(1≦i≦k)の共通集合の、ある部分集合としてもとめることができる。単語・文書対応ＤＢ３０１４に単語の文書内の位置情報まで記録されていない場合は、近似として、すべてのD(wni)(1≦i≦k)の共通集合を取ってD(W)とする。４０１４は、共起単語ベクトル計算モジュールである。再び単語・文書対応ＤＢ３０１４により、D(W)内の各単語の頻度を計数し、各単語wi がD(W)に出現する頻度ki を求める。４０１５は、分布間距離計算モジュールである。(数１)と、４０１１、４０１４で求めた単語頻度を用いて、背景単語分布と、Wを含む全文書D(W)内の単語分布との距離 Dist{PD(W)、 P0 }を求める。４０１６は、分布間距離正規化モジュールである。D(W)中の単語数を#D(W)として、４０１２でもとめたB(・)を用いて、B(#D(W))をもとめ、Rep(W) = Dist{PD(W)、 P0 }/B(#D(W)) により、Wのrepresentativenessを求める。４０１７は、ランダムサンプリングモジュールである。４０１３で、D(W)に含まれる文書数が多すぎる場合、あらかじめ定めた数（共有データ記憶領域３０１７中に記録）を越える場合、あらかじめ定めた数の文書を選ぶために用いられる。この例では、文書数をパラメータとしているが、望ましい単語数をパラーメータとし、その適当な近傍に単語数が収まるように文書をランダムサンプリングするように設定することも可能である。

【００２８】

図５は、本発明を文献検索支援のための検索内容表示に応用する場合の構成例である。本図は、(文献1)の文献検索支援方法図１において示された構成図に沿って、ナヴィゲーションウィンドウにおける特徴語表示に本発明を適用する場合の検索装置の構成を示したものである。(文献1)の文献検索支援方法と異なるのは、特徴語表示手段ルーチン５４４において、５４４５のreprsentativenessチェックルーチンが加わること、および、５４４１特徴語抽出ルーチン、５４４２共起関係解析ルーチン、５４４３グラフ配置ルーチン、５４４４グラフ表示ルーチンにおいて、５４４１特徴語抽出ルーチン、５４４３グラフ配置ルーチン、５４４４グラフ表示ルーチンにおいて、５４４５のreprsentativenessチェックルーチンを参照することである。 reprsentativenessチェックルーチンは、全文書集合における各語のreprsentativenessについて、要求に応じてその値を返すルーチンである。各語のreprsentativenessは、あらかじめ図４で示したプログラムにより計算しておくことが可能である。

【００２９】

ユーザがキーボード５１１より検索キーワードを入力すると、検索インタフェイス５２１には、そのキーワードを含む文書の見出し等が表示され、５２２特徴語表示手段には、検索結果となる文書集合から選ばれた特徴語が表示される。５４４１の特徴語抽出ルーチンにおいて(文献1)の方法で単語をまず選出する。この中には、先にのべたごとく、「する」や「この」のような一般語が混入しているが、例えば、その単語が出現する文書集合中の単語数が、例えば１０、０００語を超えるような単語について、５４４５のreprsentativenessチェックルーチンによりreprsentativenessを調べ、その値があらかじめ定めた値（例えば1。2）より低い語を排除することにより、高頻度不要語の表示を抑制できる。さらに、５４４３グラフ配置ルーチン、５４４４グラフ表示ルーチンで５４４５のreprsentativenessチェックルーチンを参照することにより、(文献1)の方法で定める各頻度クラスの中で、「語の表示が重なる場合にreprsentativeness が高い語ほど手前に表示する」、「reprsentativeness の高い語ほど文字色を濃く表示する」等の操作を行うことは容易であるから、 reprsentativeness の高い語をよ

り目立たせるような手段で表示し、ユーザインタフェイスを改善することが可能である。さらに上では、各語のreprsentativenessを、あらかじめ図４で示したプログラムにより計算しておくこととしたが、入力キーワードごとに得られる検索結果文書の集合に対して、これを改めて全文書集合と考え、図４に示したプログラムによって、検索結果文書に含まれる各語に対し、その場でreprsentativenessを計算することも可能である。５４４５のreprsentativenessチェックルーチンをそのように設計した場合、同じ語でもキーワードごとにreprsentativenessが異なってくるため、より適切に、状況を反映した特徴語を表示することが可能である。

【００３０】

図６は、単語の自動抽出に本発明を用いる場合の構成例である。６０１は記憶装置であり、ハードディスク等を用いて文書データ、各種のプログラム等を格納する。また、プログラムの作業用領域としても利用される。以下、６０１１は、文書データ。以下の例では日本語を用いるが、言語にはよらない。６０１２は、形態素解析プログラムで、文書を構成する単語を同定する。日本語の場合は分かち書き＋品詞付け、英語の場合は原型還元等の処理を行う。この手法については特定しない。両言語とも、商用・研究用をとわずさまざまなシステムが公開されている。６０１３は、単語・文書対応付けプログラム。形態素解析の結果から、どの単語がどの文書に何回あらわれているか、逆にどの文書にどのような単語が何回あらわれているかを調べる。基本的には単語と文書をそれぞれ行・列とする行列の要素を計数により埋める作業であり、この手法については特定しない。６０１４は、単語・文書対応データベースＤＢ。上記で計算された単語・文書対応データを記録するＤＢ。６０１５は、抽出単語格納ＤＢ。６０１７は、representativeness 計算プログラム（詳細は図４）６０１８は、計算されたタームのrepresentativenessを記録するＤＢ。６０１９は、複数のプログラム間で共通に参照するデータを記録する領域である。６０１Aは、最終的な抽出の候補となる単語または単語列を選び出すプログラムである。内容については特定しないが、通常、例えば「与えられた文書形態素解析した結果から、助詞、助動詞、接辞を除いた単語集合」としてよい。６０１Bは、６０１Aの選び出した候補から、文法知

識を用いて用語として不適切な語の並びを排除するフィルタである。例えば格助詞や助動詞等が先頭や末尾にくるものなどを排除する。内容については特定しないが、例えば(文献２)で紹介された論文中にいくつか例がある。６０１Bの選び出した候補は、６０１Ｃにより、特定の指標に基づき重要度を計算し、それがあらかじめ定めた値より低いものを排除したり、重要度に従ってソートして出力する。ここでは、もっとも頻繁に用いられる指標の名前にしたがって、tf_idfフィルタプログラムと呼ぶが、実際に用いる指標は、 tf_idf以外の任意の指標であってよい。６０１６は、作業用の領域である。６０２は、入力装置、６０３は、通信装置、６０４は、メインメモリ、６０５は、CPU、６０６は、ディスプレイ、キーボード等より構成される端末装置、である。一般の単語抽出方法においては、６０１７、６０１８は用いられない。６０１Cの出力に対して、６０１７、６０１８により各候補のrepresentativenessを参照し、あらかじめ定めた値（例えば1。2）よりその値が小さいものを排除する。さらに変形として、６０１Cで６０１７、６０１８を用いて直接各候補のrepresentativenessを参照し、 representativenessのみを用いて用語候補の選別を行うことも考えられる。

【００３１】

図６に述べた構成の単語の自動抽出方法を用いて、人工知能関係の１８７０本の論文の要旨から用語抽出を行う実験を行ったところ、６０１Ａ、６０１Ｂにより約１８０００個の用語候補が抽出された。６０１Ｃで、representativenessのみを用いた場合と、まずtf-idfを用いて用語候補をソートし、その出力に対してrepresentativenessを用いて非重要語の除去を行う二つの場合を行ったところ、最終的に抽出された用語候補は両者とも等しく約５０００語であるが、後者の場合の方が、頻度順に近い順序で用語が抽出される傾向にあるため、人間に提示して最終判断を仰ぐ場合は、見なれた語が先頭に近く現れる後者の方が、有る意味で自然であるともいえる。

【００３２】

【発明の効果】

本発明で提案する representativeness を用いる事により、文書集合中のタームについて、(1) 数学的な意味付けが明瞭であり、(2) 高頻度タームと低頻度タ

ームの比較が自然にできる。(3) 閾値の設定が自然にできる。(4) 任意の長さのタームに対して適用できる。ようなタームの representativeness 計算方法、すなわち、単語または単語列の重要度を計算する方法が実現でき、単語情報検索インタフェース、単語抽出システム等の精度の向上に役立てることができる。

【図面の簡単な説明】

【図１】

特徴単語提示ウィンドウを持つ情報検索支援システムの例。

【図２】

二つの単語分布の間の距離を示す例。

【図３】

提案する単語の重要度計算方法を実現するための装置構成。

【図４】

representativeness 計算プログラムの構成。

【図５】

文献検索支援のための検索内容表示にrepresentativenessを応用する場合の構成例。

【図６】

単語の自動抽出にrepresentativenessを応用する場合の構成例。

【図７】

提案する単語の重要度が、検索結果の要約にふさわしいと判断される単語の優先順位を高める力を他の指標と比較する実験結果のグラフ。

【図８】

提案する単語の重要度が、検索結果の要約にふさわしくないと判断される単語の優先順位を下げる力を他の指標と比較する実験結果のグラフ。

【符号の説明】

３０１：記憶装置

３０１１：文書データ３

０１２：形態素解析プログラム

３０１３：単語・文書対応付けプログラム

３０１４：単語・文書対応データベース（ＤＢ）

３０１５：representativeness計算プログラム

３０１６：representativenessＤＢ

３０１７：共有データ記録領域

３０１８：作業用の領域

３０２：入力装置

３０３：通信装置

３０４：メインメモリ、

３０５：CPU

３０６：端末装置、

４０１１：背景単語分布計算モジュール

４０１２：ベースライン関数推定モジュール

４０１３：部分文書集合抽出モジュール

４０１４：共起単語ベクトル計算モジュール

４０１５：分布間距離計算モジュール

４０１６：分布間距離正規化モジュール

４０１７：ランダムサンプリングモジュール

５４４：特徴単語表示手段作動ルーチン

５４４１：特徴語抽出ルーチン

５４４２：共起関係解析ルーチン

５４４３：グラフ配置ルーチン

５４４４：グラフ表示ルーチン

６０１：記憶装置

６０１１：文書データ。

６０１２：形態素解析プログラム

６０１３：単語・文書対応付けプログラム

６０１４：単語・文書対応データベースＤＢ

６０１５：抽出単語格納ＤＢ。

６０１６：作業用の領域

６０１７：representativeness 計算プログラム

６０１８：representativenessＤＢ。

６０１９：共通データ記録領域

６０１A：候補単語列抽出プログラム

６０１B：文法フィルタ

６０１C：フィルタプログラム

６０２：入力装置、

６０３：通信装置、

６０４：メインメモリ、

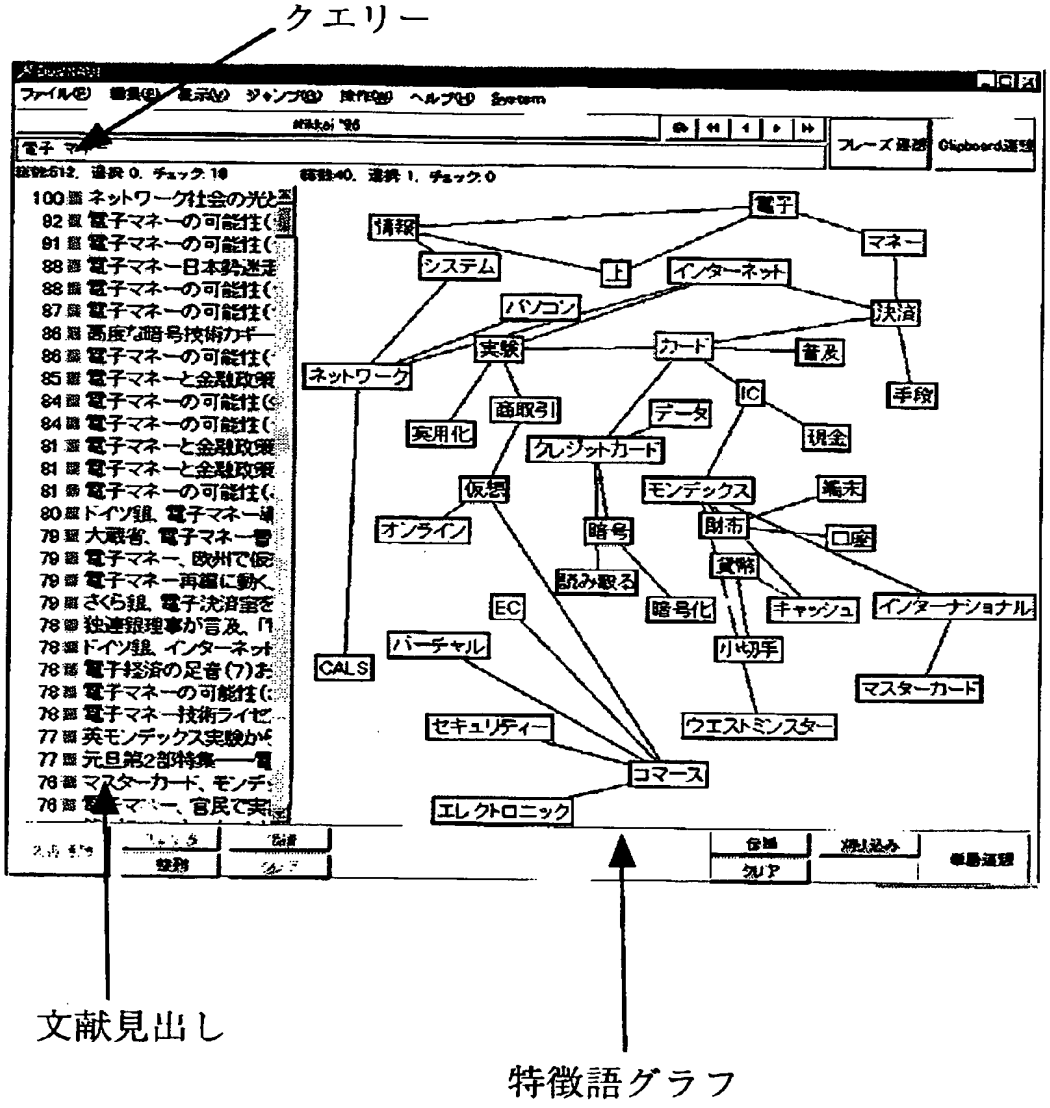
６０５：CPU

６０６：ディスプレイ、キーボード等より構成される端末装置。

【書類名】　図面

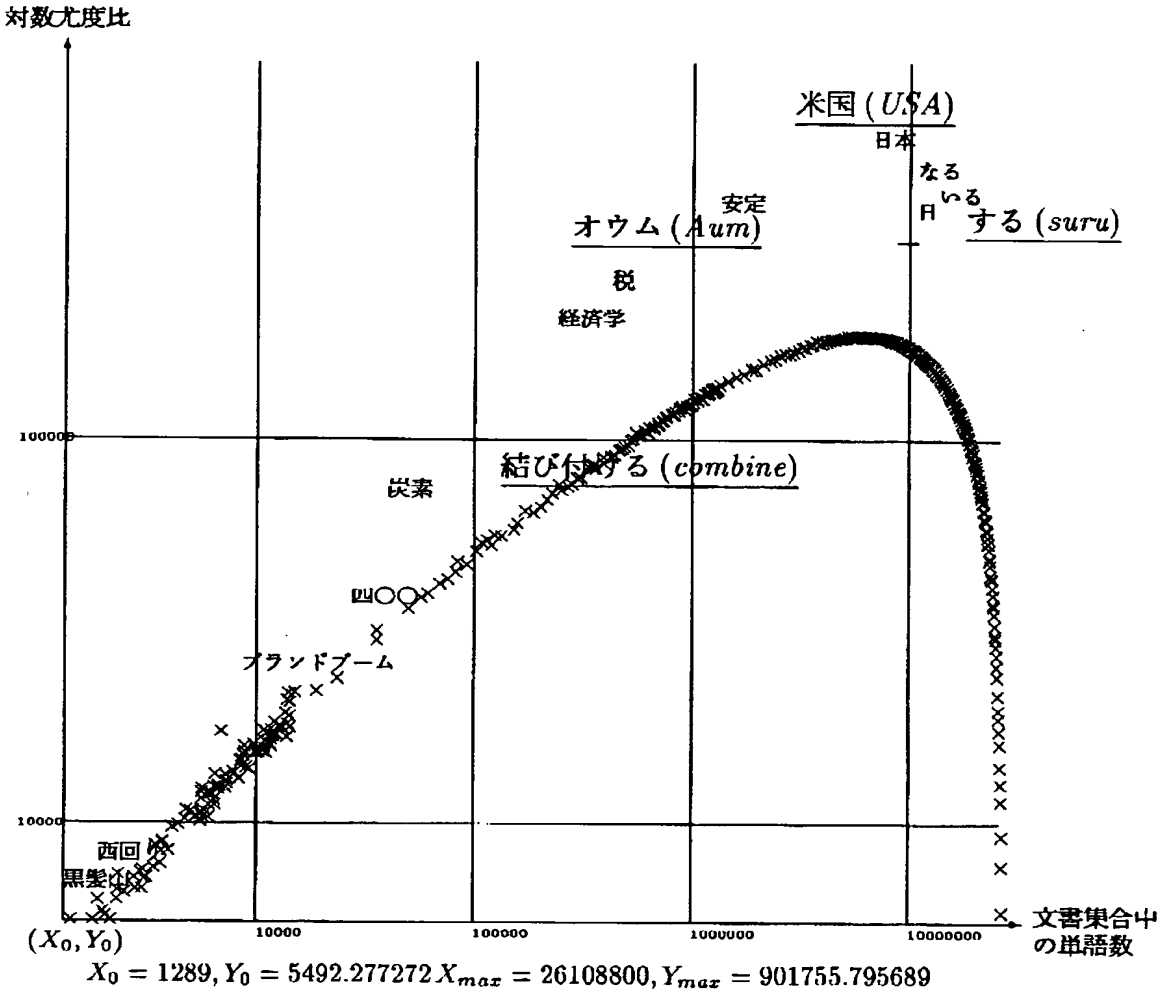
【図１】

図 1


クエリー
電子
情報
マネー
システム
上
インターネット
パソコン
決済
実験
カード
普及
ネットワーク
IC
手段
函取引
データ
実用化
現金
クレジットカード
仮想
モンデックス
端末
オンライン
暗号
財布
口座
貨幣
読み取る
暗号化
キャッシュ
インターナショナル
EC
小切手
バーチャル
CALS
マスターカード
セキュリティー
ウエストミンスター
コマース
エレクトロニック
文献見出し
特徴語グラフ

【図２】

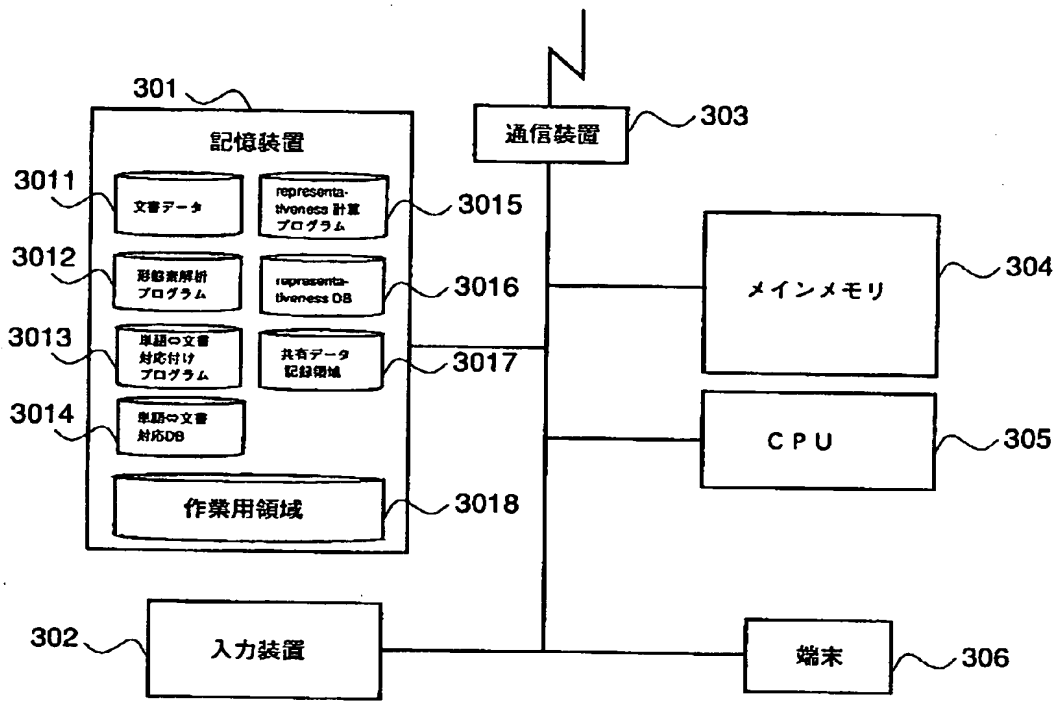
図２



$X_0 = 1289, Y_0 = 5492.277272 X_{max} = 26108800, Y_{max} = 901755.795689$

【図３】

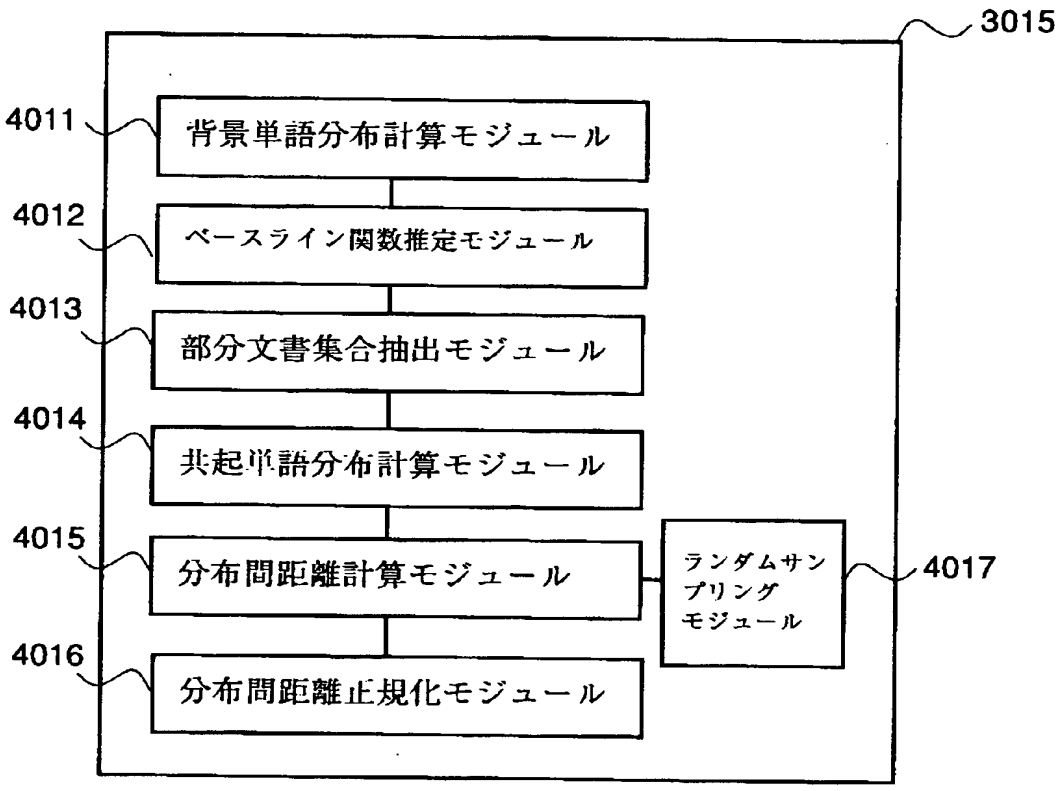
図３


301
記憶装置
3011
文書データ
representa-tiveness 計算プログラム
3015
3012
形態素解析プログラム
representa-tiveness DB
3016
3013
単語⇔文書対応付けプログラム
共有データ記録領域
3017
3014
単語⇔文書対応DB
作業用領域
3018
通信装置
303
メインメモリ
304
CPU
305
302
入力装置
端末
306

【図４】

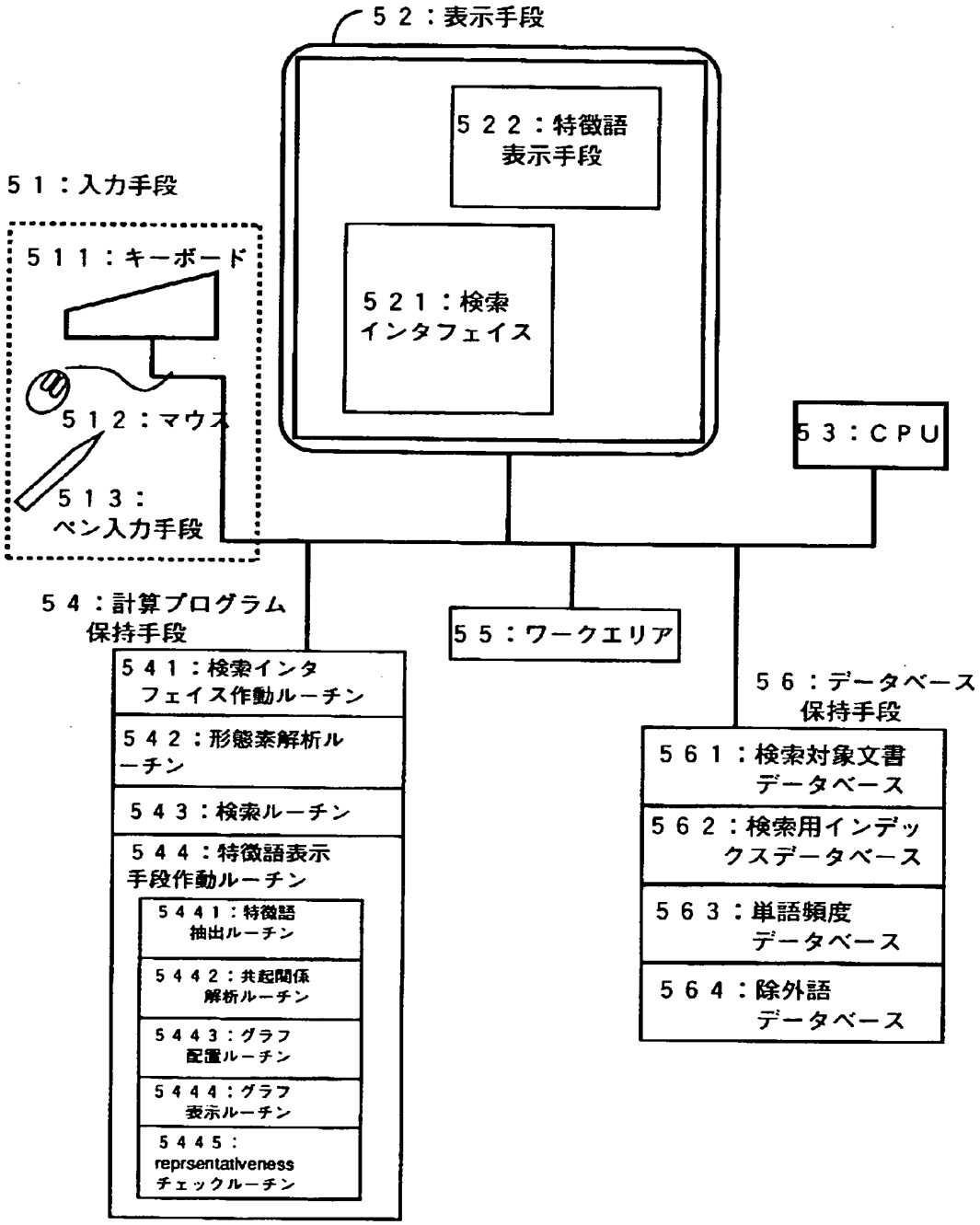
図４

representativeness 計算プログラム



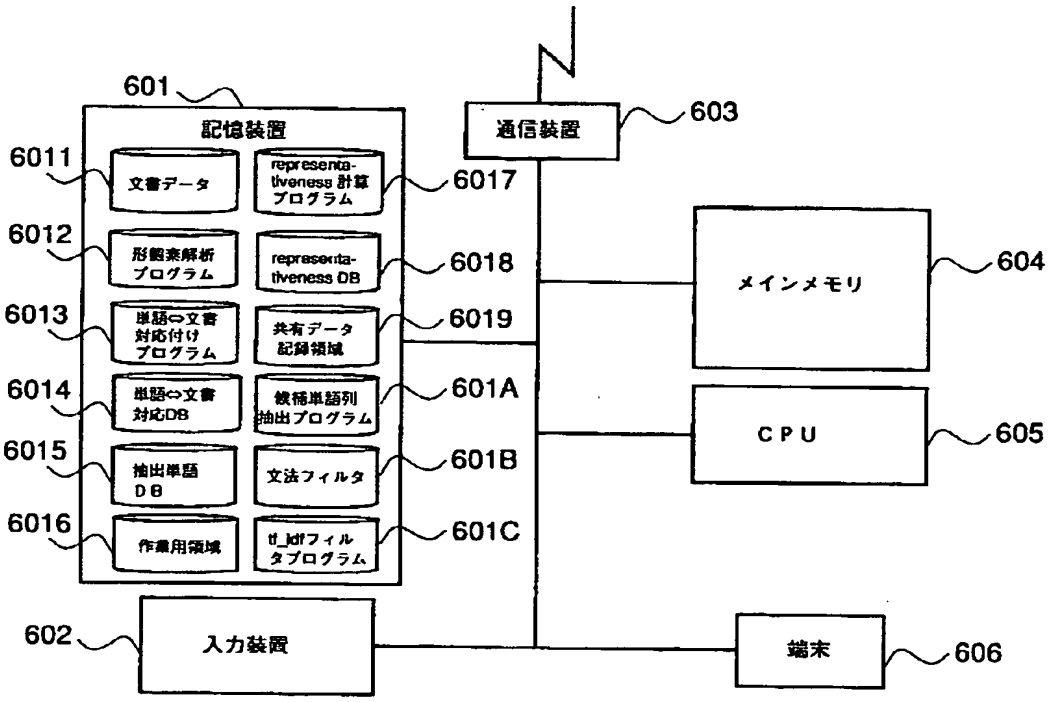
【図５】

図５


５２：表示手段
５２２：特徴語
表示手段
５２１：検索
インタフェイス
５１：入力手段
５１１：キーボード
５１２：マウス
５１３：
ペン入力手段
５３：ＣＰＵ
５４：計算プログラム
保持手段
５５：ワークエリア
５６：データベース
保持手段
５４１：検索インタ
フェイス作動ルーチン
５４２：形態素解析ル
ーチン
５４３：検索ルーチン
５４４：特徴語表示
手段作動ルーチン
５４４１：特徴語
抽出ルーチン
５４４２：共起関係
解析ルーチン
５４４３：グラフ
配置ルーチン
５４４４：グラフ
表示ルーチン
５４４５：
reprsentativeness
チェックルーチン
５６１：検索対象文書
データベース
５６２：検索用インデッ
クスデータベース
５６３：単語頻度
データベース
５６４：除外語
データベース

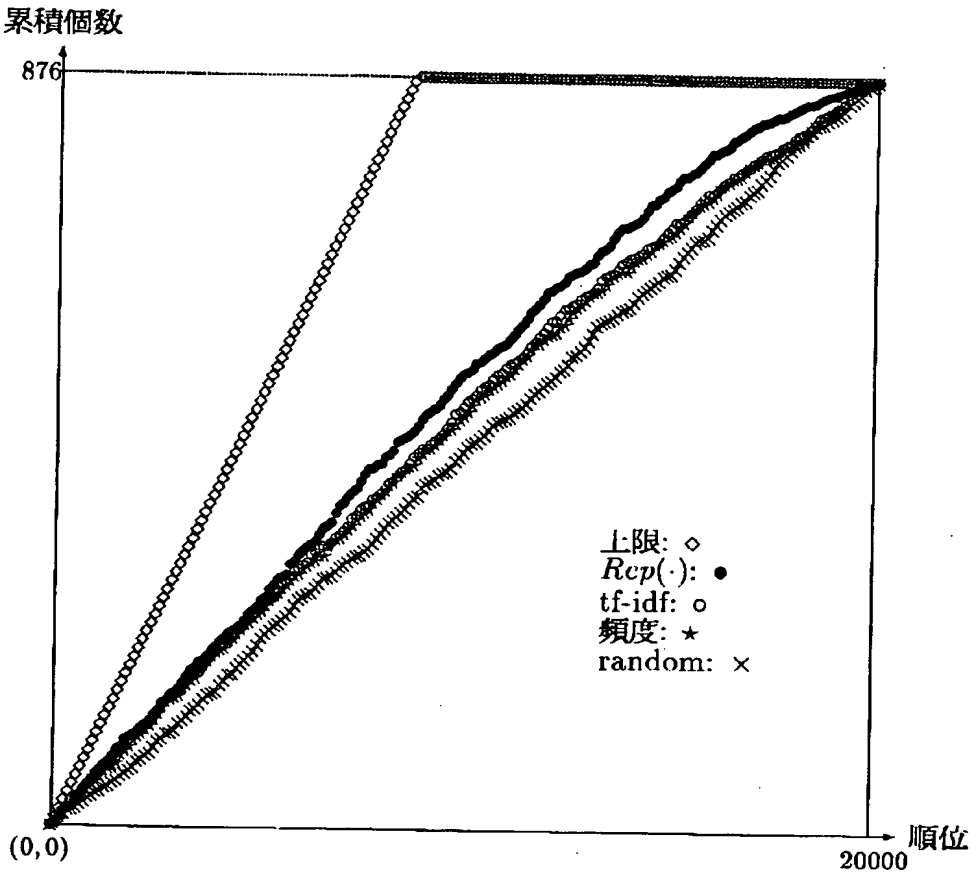
【図６】

図６


601
記憶装置
6011
文書データ
representa-tiveness 計算プログラム
6017
6012
形態素解析プログラム
representa-tiveness DB
6018
6013
単語⇔文書対応付けプログラム
共有データ記録領域
6019
6014
単語⇔文書対応DB
候補単語列抽出プログラム
601A
6015
抽出単語DB
文法フィルタ
601B
6016
作業用領域
tf_idfフィルタプログラム
601C
602
入力装置
通信装置
603
メインメモリ
604
CPU
605
端末
606

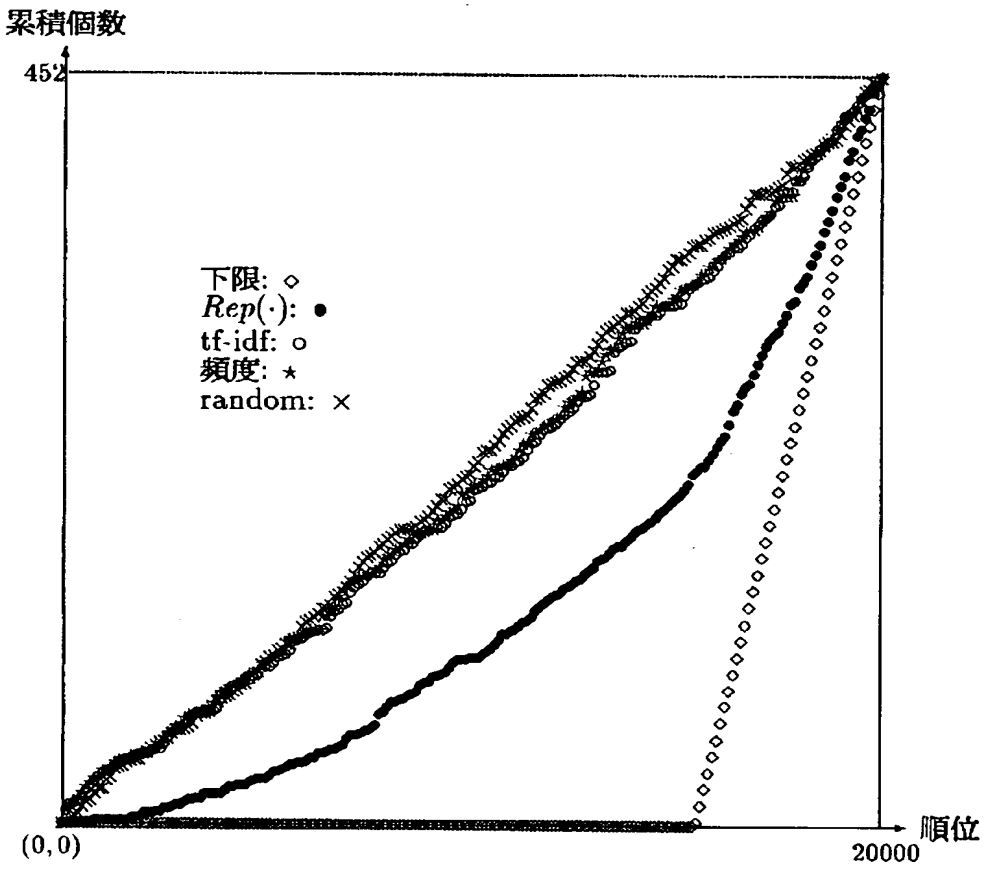
【図７】

図７



【図８】

図８



【書類名】　要約書

【要約】

【課題】　情報検索において重要である単語（列）を選択する方法においては、従来、高頻度一般語の除去ができない、重要／非重要を分ける閾値の設定が恣意的になりがちであるなど、いくつかの問題があった。本発明の目的は、このような問題の無い方法を提案し、情報検索支援の精度を向上させることである。

【解決手段】　上記の問題を、抽出すべき単語を含む部分文書集合中の単語分布と、もととなる文書集合中の単語分布との乖離度を、該部分文書集合中の単語数で正規化することにより解決する。

【選択図】　図３

出　願　人　履　歴　情　報

識別番号　　〔０００００５１０８〕

１．変更年月日　　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所